# The use of measures of influence in epidemiology

Helfenstein, Ulrich ; Minder, Christoph

Abstract: In epidemiological studies the units of observation often consist of political entities such as countries, each of which has its own specific inner structure. When a multiple regression is performed it is therefore of particular interest to analyse not only the overall behaviour of the dataset, but in addition, to investigate how each individual country contributes to, and deviates from, this overall behaviour. By means of the example 'relation between infant mortality and structural data of countries' several ways are discussed of how each individual country can influence the regression model. Firstly the potential influence which each country might exhibit due to the explanatory variables alone is analysed. Then the actual influence of each country is analysed by taking the explanatory variables and the target variable into account simultaneously. This is done by means of statistical measures not generally familiar to epidemiologists, which have been developed in recent years (leverage values, Cook's distances). These measures also point to deviations of countries from the model, and suggest directions in which to search for explanation. Finally the influence of the 'size' of the countries is investigated.

DOI: https://doi.org/10.1093/ije/19.1.197

International Journal of Epidemiology
Vol. 19, No. 1

Printed in Great Britain

# The Use of Measures of Influence in Epidemiology

ULRICH HELFENSTEIN* AND CHRISTOPH MINDER**

Helfenstein U (Biostatistical Centre for the Medical Department, University of Zurich, Plattenstrasse 54, CH-8032, Zurich, Switzerland) and Minder C. *International Journal of Epidemiology* 1990, **19**: 197–204.
In epidemiological studies the units of observation often consist of political entities such as countries, each of which has its own specific inner structure. When a multiple regression is performed it is therefore of particular interest to analyse not only the overall behaviour of the dataset, but in addition, to investigate how each individual country contributes to, and deviates from, this overall behaviour.

By means of the example 'relation between infant mortality and structural data of countries' several ways are discussed of how each individual country can influence the regression model. Firstly the *potential* influence which each country might exhibit due to the explanatory variables alone is analysed. Then the *actual* influence of each country is analysed by taking the explanatory variables and the target variable into account simultaneously. This is done by means of statistical measures not generally familiar to epidemiologists, which have been developed in recent years (leverage values, Cook's distances). These measures also point to deviations of countries from the model, and suggest directions in which to search for explanation. Finally the influence of the 'size' of the countries is investigated.

When a multiple regression is performed to analyse data, results and interpretations are usually based on summary statistics such as slopes, coefficient of determination and others.[1] In epidemiological studies, however, the units of observation often consist of countries or other political entities, each of which has its own specific inner structure. When a multiple regression is performed it is therefore of particular interest to analyse not only the overall behaviour of the dataset, but in addition, to investigate how each individual country contributes to, and deviates from, this overall behaviour. Similar deliberations apply to ecological or surveillance studies where the units of observation consist of groups of people such as occupations, social classes, communities, etc. Subsequently, an example is presented to illustrate some of the concepts underlying measures of influence: The relation between infant mortality and structural data of countries.

Several ways are described how each individual country may influence the regression model and therefore the conclusions about the relation between infant mortality and structural data. Firstly, we consider the *potential* influence of each country due to its values of the explanatory structural variables alone. Then the *actual* influence of each country is analysed by taking its values for explanatory variables and target variable simultaneously into account. This is done by means of *case statistics*, statistical measures which have been developed during the last years.[2,3] They are presented in the next section. These analyses will also point to deviations of countries from the model, and suggest directions in which to search for explanation.

A further particularity of this kind of data is that the units of observation may differ strongly in 'size'. Approximately one-fifth of the world population lives in China. It seems therefore at first sight evident that China should receive a much larger weight than a 'small' country. This conclusion is, as will be shown, however, doubtful.

Subsequently the relation between infant mortality and structural data of 125 countries with more than one million inhabitants each is investigated. The data stem from the UN, the world bank and the OECD. A larger set of structural data may be found in the 'Weltalmanach' 1986.[4] In our example, infant mortality is the target variable and the explanatory variables are: number of inhabitants, density of population, gross domestic product per capita ($), food supply in per cent, imported area cultivated (%), number of inhabitants per physician and illiteracy (%). These data describe different characteristics of a country, like economic situation, educational standard of the population, medical supply etc. The regression is used to investi-

*Biostatistical Center of the Medical Department, University of Zurich Plattenstrasse 54, CH-8032 Zurich, Switzerland.
**Institut für Sozial- und Präventivmedizin, Finkenhubelweg 11, CH-3012 Bern, Switzerland.

gate which combination of these structural variables can best 'explain' infant mortality. For ease of reference the data[4] are presented in Table 1.

Several investigations about relations between infant mortality and explanatory variables have been published, all using overall statistics which arise from regression models. Since the main concern of the present study is the use of influence measures, we refer to the article of Woodhandler and Himmelstein[5] for a review of past work on these relationships.

## STATISTICAL METHODS

Firstly, the different sources of the influence which the individual countries exert on the regression model are explained then the precise mathematical formulae are given, with graphs of the particular example given to assist interpretation.

### *Potential Influence*

One aspect of influence is determined by the values of the explanatory variables. The case statistics describing this aspect are called leverage values.[6] The closer the values of the regressor variables lie to the border of the observed region, the larger are the corresponding leverage values (compare Figure 3 and its description in the next sections). Since the values of the target variable do not enter into these statistics, it is possible that a case with a large leverage value turns out to have no marked influence on the model. To emphasize this aspect, Cook and Weisberg[7] call them *potential values*.

In order to give a mathematical formula to this concept, assume that the relation between the target variable and the p explanatory variables is represented by the regression model:

(1) $y = X\beta + e$

$y$ is the vector of observed responses, X is the matrix of explanatory data and $e$ is a random vector with mean $O$ and covariance matrix $\sigma^2 1$. The vector of fitted responses $\hat{y}$ may be obtained from $y$ by the linear operation:

(2) $\hat{y} = Hy$,

where $H = X(X^TX)^{-1}X^T$ is called the 'hat' matrix because it transforms the vector $y$ into the vector of fitted responses $\hat{y}$. The diagonal elements $h_{ii}$ of the hat matrix H are called *potential values* (or leverage values).[6]

A helpful representation of $h_{ii}$ is given by:

(3) $h_{ii} = 1/n + (x_i - \bar{x})^T S_x^{-1} (x_i - \bar{x}) / (n-1)$,

where $S_x$ is the covariance matrix of the explanatory variables. This case statistic has a useful geometric interpretation: If the term $1/n$ on the right side is dropped, remainder is proportional to the Mahalanobis distance from $x_i$ to the centre $\bar{x}$. Points lying on elliptical contours have the same Mahalanobis distance from the centre and therefore the same potential influence on the regression model.

### *Actual Influence*

*Cook's distance.* A further aspect of influence is determined by the residuals ie by the deviations of the observed values of the target variable (infant mortality) from the fitted values. Potential values and residuals are combined into a single case statistic called Cook's distance.[6] This measure contains information from the explanatory variables and from the target variable and it determines the *actual* influence of each country on the model.

In mathematical terms Cook's distance of the i-th country is given by[6]:

(4) $D_i = (1/p)\, r_i^2\, (h_{ii}/(1-h_{ii}))$,

where $r_i = e_i \sigma_i^{-1} (1-h_{ii})^{-1/2}$

$D_i$ is essentially composed of two parts: The first is the square of the studentized residual $r_i$, ie a measure of the discrepancy between the observed and the fitted value corrected for its individual precision. The second part is a monotonic increasing function of the i-th potential value $h_{ii}$. Thus, a large value of $D_i$ may be due to large $r_i$, large $h_{ii}$, or both.

*Size of the units.* Different approaches have been suggested to solve the problem of 'size' (different number of inhabitants, infants, etc). In their investigation of cardiovascular mortality rates in 161 local authorities in England, Fryer *et al.*[8] eg proposed the use of weights inversely proportional to binomial variance. Pocock *et al*[9] performed a thorough statistical analysis of the problem. They found that the variation of mortality rates between political units is composed of three components:

(i) explained variation;
(ii) unexplained variation;
(iii) binomial sampling variation.

The explained variation (i) is of interest because it can contribute to a better understanding of a disease process and its possible causes. Since one can not expect to include all explanatory variables, the component (ii) is present. In each country the number of infants may be thought of as being a sample of a hypothetical population with an unknown 'true' mortality rate. This leads to component (iii); (ii) and (iii) together give the variation about regression.

The 'size' of the countries or the number of infants are only of concern with regard to the binomial variation (iii). If the binomial variations are small, the unexplained component (ii) dominates, and a weighted regression may lead to an overweighting of the 'large' countries and thus distort the results. If the unex-

TABLE 1 *The data: Structural data of countries.*

| Country | Area | Inhabitants | GDP | Cult. area | Inh./phys. | Inf. mort. | Illiteracy | Birth rate | Food sup. |
|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 647.0 | 14.5 | 221 | 12.0 | 16730 | 20.5 | 80.0 | 5.40 | 14 |
| Albania | 29.0 | 2.8 | 535 | 24.0 | 960 | 4.4 | — | 2.80 | — |
| Algeria | 2382.0 | 20.5 | 2400 | 3.0 | 2630 | 11.1 | 65.0 | 4.70 | 21 |
| Angola | 1247.0 | 8.3 | 990 | 1.0 | 14910 | 16.5 | 95.0 | 4.90 | — |
| Argentina | 2777.0 | 29.6 | 2030 | 13.0 | 430 | 4.4 | 7.0 | 2.50 | 5 |
| Australia | 7686.0 | 15.4 | 10780 | 6.0 | 560 | 1.0 | 0 | 1.60 | 5 |
| Austria | 84.0 | 7.6 | 9210 | 20.0 | 400 | 1.3 | 1.0 | 1.30 | 7 |
| Bangladesh | 144.0 | 94.6 | 130 | 68.0 | 10940 | 13.3 | 74.0 | 4.70 | 20 |
| Belgium | 31.0 | 9.9 | 9160 | 27.0 | 400 | 1.2 | 1.0 | 1.20 | 12 |
| Benin | 113.0 | 3.7 | 290 | 16.0 | 16980 | 11.7 | 72.0 | 4.90 | 17 |
| Bhutan | 47.0 | 1.4 | 114 | 5.0 | 18160 | 16.3 | 80.0 | 4.30 | — |
| Bolivia | 1099.0 | 6.1 | 510 | 3.0 | 3830 | 12.6 | 37.0 | 4.30 | — |
| Brazil | 8512.0 | 129.7 | 1890 | 5.0 | 2210 | 7.3 | 24.0 | 3.10 | 9 |
| Bulgaria | 111.0 | 8.9 | 4500 | 39.0 | 410 | 2.0 | 9.0 | 1.50 | — |
| Burkina Faso | 274.0 | 6.6 | 180 | 10.0 | 32767 | 15.7 | 95.0 | 4.80 | 25 |
| Burma | 676.0 | 35.3 | 180 | 15.0 | 4660 | 9.6 | 34.0 | 3.80 | 14 |
| Burundi | 28.0 | 4.4 | 240 | 50.0 | 32767 | 12.3 | 75.0 | 4.70 | — |
| Cambodia | 181.0 | 6.9 | 113 | 17.0 | 32767 | 14.6 | 64.0 | 4.50 | — |
| Cameroon | 475.0 | 9.2 | 800 | 16.0 | 13990 | 9.2 | 81.0 | 4.60 | 9 |
| Canada | 9976.0 | 25.0 | 12000 | 5.0 | 550 | 1.0 | 1.0 | 1.50 | 7 |
| Central African R. | 623.0 | 2.5 | 280 | 5.0 | 26430 | 11.9 | 67.0 | 4.10 | 21 |
| Chad | 1284.0 | 4.8 | 80 | 3.0 | 32767 | 16.1 | 85.0 | 4.20 | 19 |
| Chile | 757.0 | 11.7 | 1870 | 8.0 | 1930 | 2.7 | 16.0 | 2.30 | 15 |
| China | 9561.0 | 1024.0 | 290 | 15.0 | 1810 | 6.7 | 32.0 | 1.90 | 16 |
| Colombia | 1139.0 | 27.7 | 1410 | 9.0 | 1710 | 5.4 | 19.0 | 2.90 | 10 |
| Costa Rica | 51.0 | 2.4 | 1020 | 10.0 | 1460 | 1.8 | 10.0 | 3.00 | 9 |
| Cuba | 115.0 | 9.9 | 800 | 28.0 | 710 | 1.7 | 5.0 | 1.60 | — |
| Czechoslovakia | 128.0 | 15.4 | 5970 | 42.0 | 360 | 1.6 | 5.0 | 1.50 | 10 |
| Denmark | 43.0 | 5.1 | 11490 | 63.0 | 480 | 0.8 | 1.0 | 0.99 | 12 |
| Dominican Rep. | 49.0 | 6.0 | 1380 | 25.0 | 2320 | 6.5 | 30.0 | 3.40 | 18 |
| East Germany | 108.0 | 16.7 | 8600 | 46.0 | 520 | 1.2 | — | 1.50 | — |
| Ecuador | 284.0 | 9.3 | 1430 | 9.0 | 760 | 7.8 | 19.0 | 3.70 | 9 |
| Eqypt | 1001.0 | 45.9 | 700 | 3.0 | 970 | 10.4 | 56.0 | 3.50 | 34 |
| El Salvador | 21.0 | 5.2 | 710 | 34.0 | 3220 | 7.2 | 38.0 | 4.00 | 17 |
| Ethiopia | 1222.0 | 33.6 | 140 | 12.0 | 32767 | 12.2 | 85.0 | 4.70 | 9 |
| Finland | 338.0 | 4.9 | 10440 | 8.0 | 530 | 0.7 | 0 | 1.38 | 7 |
| France | 547.0 | 55.1 | 10390 | 32.0 | 580 | 1.0 | 1.0 | 1.40 | 10 |
| Ghana | 239.0 | 12.2 | 320 | 12.0 | 7630 | 8.6 | 73.0 | 4.90 | 19 |
| Greece | 132.0 | 9.9 | 3970 | 30.0 | 420 | 1.4 | 19.0 | 1.40 | 11 |
| Guatemala | 109.0 | 7.9 | 1120 | 17.0 | 8610 | 6.6 | 68.0 | 3.80 | 6 |
| Guinea | 246.0 | 5.2 | 300 | 17.0 | 17110 | 19.0 | 80.0 | 4.90 | — |
| Haiti | 28.0 | 5.3 | 320 | 32.0 | 8200 | 11.0 | 77.0 | 3.20 | — |
| Honduras | 112.0 | 4.1 | 670 | 16.0 | 3120 | 8.3 | 40.0 | 4.40 | 10 |
| Hungary | 93.0 | 10.7 | 2150 | 58.0 | 400 | 2.0 | 1.0 | 1.20 | 9 |
| India | 3288.0 | 730.0 | 260 | 57.0 | 3690 | 9.4 | 64.0 | 3.40 | 9 |
| Indonesia | 1919.0 | 159.4 | 560 | 9.0 | 11530 | 10.2 | 38.0 | 3.40 | 11 |
| Irak | 438.0 | 14.6 | 1800 | 12.0 | 1800 | 7.3 | 82.0 | 4.50 | — |
| Iran | 1648.0 | 42.1 | 2000 | 10.0 | 6090 | 10.2 | 50.0 | 4.10 | 14 |
| Ireland | 70.0 | 3.5 | 4810 | 14.0 | 780 | 1.1 | 2.0 | 2.00 | 13 |
| Israel | 21.0 | 4.1 | 5360 | 20.0 | 370 | 1.6 | 16.0 | 2.40 | 12 |
| Italy | 301.0 | 56.8 | 6350 | 42.0 | 340 | 1.4 | 2.0 | 1.10 | 12 |
| Ivory Coast | 322.0 | 9.3 | 720 | 12.0 | 21040 | 11.9 | 65.0 | 4.80 | 20 |
| Jamaica | 11.0 | 2.3 | 1300 | 24.0 | 2830 | 1.0 | 10.0 | 2.70 | 19 |
| Japan | 372.0 | 117.2 | 10100 | 13.0 | 780 | 0.7 | 1.0 | 1.30 | 13 |
| Jordan | 98.0 | 3.3 | 1710 | 14.0 | 1700 | 6.5 | 30.0 | 4.50 | 17 |
| Kenya | 583.0 | 18.8 | 340 | 4.0 | 7890 | 7.7 | 53.0 | 5.50 | 8 |
| Kongo | 342.0 | 1.7 | 1230 | 2.0 | 5510 | 6.8 | 84.0 | 4.30 | 19 |
| Kuwait | 18.0 | 1.7 | 18180 | 0.1 | 570 | 3.2 | 40.0 | 3.50 | 14 |

TABLE 1 *Continued*

| Country | Area | Inhabitants | GDP | Cult. area | Inh./phys. | Inf. mort. | Illiteracy | Birth rate | Food sup. |
|---|---|---|---|---|---|---|---|---|---|
| Laos | 237.0 | 4.2 | 95 | 4.0 | 20060 | 15.9 | 56.0 | 4.20 | — |
| Lebanon | 10.0 | 2.6 | 1900 | 34.0 | 540 | 3.9 | 14.0 | 2.90 | — |
| Lesotho | 30.0 | 1.4 | 470 | 10.0 | 18640 | 9.4 | 48.0 | 4.20 | — |
| Liberia | 111.0 | 2.1 | 470 | 4.0 | 9610 | 9.1 | 75.0 | 5.00 | 22 |
| Libya | 1760.0 | 3.4 | 7500 | 1.0 | 730 | 9.5 | 50.0 | 4.50 | 18 |
| Madagascar | 587.0 | 9.4 | 290 | 5.0 | 10170 | 11.6 | 50.0 | 4.70 | 14 |
| Malawi | 118.0 | 6.4 | 210 | 24.0 | 32767 | 13.7 | 75.0 | 5.60 | 8 |
| Malaysia | 330.0 | 15.1 | 1870 | 20.0 | 7910 | 2.9 | 40.0 | 2.90 | 13 |
| Mali | 1240.0 | 7.5 | 150 | 2.0 | 22130 | 13.2 | 90.0 | 4.80 | 20 |
| Mauritania | 1031.0 | 1.8 | 440 | 1.0 | 14350 | 13.2 | 83.0 | 4.30 | 5 |
| Mexico | 1958.0 | 75.1 | 2240 | 12.0 | 1830 | 5.3 | 17.0 | 3.40 | 4 |
| Mongolia | 1565.0 | 1.8 | 1050 | 1.0 | 450 | 5.1 | 5.0 | 3.40 | — |
| Morocco | 459.0 | 22.1 | 750 | 18.0 | 10750 | 12.5 | 72.0 | 4.00 | 23 |
| Mozambique | 799.0 | 13.3 | 211 | 4.0 | 32767 | 10.5 | 67.0 | 4.90 | — |
| Nepal | 141.0 | 15.7 | 170 | 17.0 | 30060 | 14.5 | 81.0 | 4.30 | 4 |
| Netherlands | 42.0 | 14.4 | 9910 | 25.0 | 540 | 0.8 | 1.0 | 1.20 | 15 |
| New Zealand | 269.0 | 3.2 | 7410 | 2.0 | 650 | 1.2 | 1.0 | 1.60 | 6 |
| Nicaragua | 130.0 | 3.1 | 900 | 13.0 | 1800 | 8.6 | 10.0 | 4.50 | 18 |
| Niger | 1267.0 | 5.8 | 240 | 3.0 | 32767 | 13.2 | 90.0 | 5.20 | 23 |
| Nigeria | 924.0 | 89.0 | 760 | 33.0 | 12550 | 10.9 | 66.0 | 5.00 | 14 |
| North Korea | 121.0 | 19.2 | 1360 | 19.0 | 430 | 3.2 | 15.0 | 3.00 | — |
| Norway | 324.0 | 4.1 | 13820 | 3.0 | 520 | 0.8 | 1.0 | 1.21 | 7 |
| Oman | 212.0 | 1.1 | 6240 | 0.2 | 1900 | 12.3 | 75.0 | 4.70 | 13 |
| Pakistan | 796.0 | 92.9 | 390 | 26.0 | 3480 | 12.1 | 76.0 | 4.20 | 14 |
| Panama | 76.0 | 2.1 | 2070 | 7.0 | 980 | 3.3 | 15.0 | 2.80 | 10 |
| Papua new Guinea | 462.0 | 3.2 | 790 | 1.0 | 13590 | 9.9 | 68.0 | 3.40 | 30 |
| Paraguay | 407.0 | 3.5 | 1410 | 3.0 | 1710 | 4.5 | 16.0 | 3.10 | — |
| Peru | 1285.0 | 18.7 | 1040 | 3.0 | 1390 | 8.3 | 20.0 | 3.40 | 19 |
| Philippines | 300.0 | 52.0 | 760 | 33.0 | 7970 | 5.1 | 25.0 | 3.10 | 8 |
| Poland | 313.0 | 36.6 | 3952 | 49.0 | 570 | 2.0 | 2.0 | 1.90 | 18 |
| Portugal | 92.0 | 10.1 | 2190 | 39.0 | 540 | 2.6 | 22.0 | 1.80 | 16 |
| Romania | 238.0 | 22.7 | 2400 | 45.0 | 680 | 2.9 | 2.0 | 1.70 | 3 |
| Rwanda | 26.0 | 5.7 | 270 | 39.0 | 31510 | 12.6 | 50.0 | 5.40 | — |
| Saudi Arabia | 2150.0 | 10.4 | 12180 | 0.5 | 1670 | 10.8 | 75.0 | 4.30 | 14 |
| Senegal | 196.0 | 6.3 | 440 | 12.0 | 13800 | 15.5 | 90.0 | 4.80 | 28 |
| Sierra Leone | 72.0 | 3.5 | 380 | 25.0 | 16220 | 19.0 | 85.0 | 4.90 | 23 |
| Singapore | 0.6 | 2.5 | 6620 | 12.0 | 1150 | 1.1 | 17.0 | 1.70 | 7 |
| Somalia | 638.0 | 5.3 | 250 | 2.0 | 14290 | 18.4 | 40.0 | 4.80 | 33 |
| South Africa | 1123.0 | 26.1 | 2450 | 12.0 | 2180 | 5.5 | 43.0 | 4.00 | 4 |
| South Korea | 98.0 | 40.6 | 2010 | 23.0 | 1440 | 3.2 | 7.0 | 2.30 | 12 |
| Soviet Union | 22275.0 | 276.3 | 5500 | 10.0 | 270 | 3.3 | 0 | 1.90 | 12 |
| Spain | 505.0 | 38.2 | 4800 | 41.0 | 460 | 1.0 | 13.0 | 1.50 | 12 |
| Sri Lanka | 66.0 | 15.4 | 330 | 33.0 | 7170 | 3.2 | 15.0 | 2.70 | 19 |
| Sudan | 2506.0 | 20.4 | 400 | 5.0 | 8930 | 11.9 | 68.0 | 4.50 | 19 |
| Sweden | 450.0 | 8.3 | 12400 | 7.0 | 490 | 0.7 | 1.0 | 1.10 | 7 |
| Switzerland | 41.0 | 6.5 | 16390 | 10.0 | 410 | 0.8 | 1.0 | 1.10 | 9 |
| Syria | 185.0 | 10.4 | 1680 | 31.0 | 2270 | 5.8 | 42.0 | 4.60 | 24 |
| Tanzania | 945.0 | 20.4 | 240 | 6.0 | 17560 | 9.8 | 21.0 | 4.70 | 13 |
| Thailand | 513.0 | 49.5 | 810 | 35.0 | 7100 | 5.1 | 14.0 | 2.80 | 4 |
| Togo | 56.0 | 2.8 | 280 | 20.0 | 18100 | 12.2 | 82.0 | 4.90 | 26 |
| Trindad and Tobago | 5.0 | 1.2 | 6900 | 31.0 | 1360 | 2.6 | 5.0 | 2.90 | 13 |
| Tunisia | 164.0 | 6.9 | 1290 | 32.0 | 3690 | 6.5 | 38.0 | 3.40 | 14 |
| Turkey | 781.0 | 47.3 | 1230 | 36.0 | 1630 | 8.3 | 40.0 | 3.10 | 3 |
| Uganda | 236.0 | 14.6 | 220 | 28.0 | 26810 | 12.0 | 48.0 | 5.00 | 6 |
| United Arab E. | 84.0 | 1.2 | 21340 | 0.2 | 900 | 5.0 | 44.0 | 2.80 | 11 |
| United Kingdom | 244.0 | 55.7 | 9050 | 29.0 | 650 | 1.1 | 1.0 | 1.30 | 14 |
| United States | 9363.0 | 234.0 | 14090 | 21.0 | 520 | 1.1 | 1.0 | 1.60 | 8 |
| Urguay | 178.0 | 3.0 | 2490 | 11.0 | 540 | 3.4 | 6.0 | 1.80 | 7 |
| Venezuela | 912.0 | 15.1 | 4100 | 4.0 | 990 | 3.9 | 18.0 | 3.50 | 17 |

TABLE I *Continued*

| Country | Area | Inhabitants | GDP | Cult. area | Inh./phys. | Inf. mort. | Illiteracy | Birth rate | Food sup. |
|---|---|---|---|---|---|---|---|---|---|
| Vietnam | 333.0 | 57.2 | 180 | 18.0 | 4190 | 5.3 | 13.0 | 3.50 | — |
| West Germany | 249.0 | 61.0 | 11420 | 31.0 | 450 | 1.2 | 1.0 | 1.00 | 12 |
| Yemen (Aden) | 333.0 | 2.1 | 510 | 0.6 | 7200 | 14.0 | 60.0 | 4.80 | — |
| Yemen (Sana) | 195.0 | 6.2 | 510 | 14.0 | 11670 | 16.3 | 79.0 | 4.80 | 28 |
| Yugoslavia | 256.0 | 22.9 | 2570 | 31.0 | 550 | 3.4 | 15.0 | 1.50 | 6 |
| Zaire | 2345.0 | 31.2 | 160 | 3.0 | 14780 | 10.6 | 45.0 | 4.60 | — |
| Zambia | 753.0 | 6.2 | 580 | 7.0 | 7670 | 10.5 | 56.0 | 5.00 | — |
| Zimbabwe | 391.0 | 7.7 | 740 | 8.0 | 6580 | 8.3 | 21.0 | 5.40 | — |

Abbreviations are explained in the text. A dash indicates a missing value.

plained variation (ii) disappears, weighted regression with weights inversely proportional to the binomial variance needs to be applied.

In order to give a precise formula to the above concepts, denote the number of infants in the country with index i (i=1, . . ., 125) by $n_i$. During a certain time interval $d_i$ of the infants died. The observed mortality rate is $y_i = d_i/n_i$. The 'true' unknown death rate is $\pi_i$.

For the observed rate $y_i$ we may then write:

(5) $y_i = \pi_i + \xi_i$.

$\xi_i$ has a shifted binomial distribution with expectation 0 and variance $\tau_i^2 = \pi_1(1-\pi_i)/n_i$ (when $n_i$ is large $\xi_i$ is approximately normally distributed). Now, the $\pi_i$ are related to the explanatory variables by the usual multiple linear regression:

(6) $\pi_i = \beta_0 + \beta_1 \times_{1i} + \ldots . + \beta_{pi} \times_{pi} + \varepsilon_i$.

The $\varepsilon_i$ describe the unexplained part of the $\pi_i$. They have the constant variance $\sigma^2$. The p regressor variables describe the explained part. Combining (1) and (2) we get:

(7) $y_i = \beta_0 + \beta_1 \times_{1i} + \ldots . + \beta_{pi} \times_{pi} + \eta_i$,

$\eta_i$ is a random variable with expectation 0 and variance

(8) $\sigma^2 + \pi_i(1-\pi_i) / n_i$.

As one can see, the number of infants $n_i$ only affects the $\tau_i^2 = \pi_i(1-\pi_i) /n_i$. If the $\tau_i^2$ are small compared to the unexplained component $\sigma^2$, the latter dominates the regression error. In this situation, binomial weighting may lead to an overweighting of the 'large' countries. If on the other hand $\sigma^2 = 0$, the correct solution is given by the regression with weights proportional to $\tau_i^{-2}$.

Pocock *et al*[9] have proposed a maximum likelihood method, designed to estimate error components and parameters.

## RESULTS

In order to explain infant mortality rates, the following structural variables were used (see Table 1): Number of inhabitants, density of population (inhabitants per area), gross domestic product per capita ($), food supplies in per cent of imports, area cultivated (%), number of inhabitants per physician and illiteracy (% of population). Some of these variables, eg the gross domestic product, have a skewed distribution and hence were transformed logarithmically ('symmetrization').

First a stepwise multiple regression was performed without weighting the countries. From all available explanatory variables the regression showed up only two as significant: illiteracy (ILLIT) and gross domestic product per capita (GDP). The relation between each of these variables and infant mortality is presented in Figures 1 and 2.

In a second step the procedure proposed by Pocock *et al*[9] was applied to estimate the binomial component of the variance about regression (compare last section). This binomial part was found to be only 0.64‰. The variation about regression is therefore dominated by the unexplained component: a regression with Pocock's method gives practically the same results as the unweighted regression.

With the two explanatory variables ILLIT and GDP, both methods gave the same coefficient of determination $R^2 = 0.82$. Additional explanatory variables did not increase $R^2$ significantly.

The regression with binomial weights leads to a different result. The effect of weighting is illustrated in Figure 2. The figure shows a marked negative correlation between infant mortality and GDP. In this picture the individual countries are not represented by points as usual, but rather by circles. The areas of the circles are proportional to the binomial weights of the corresponding countries. The two straight lines are calculated from simple regressions with and without weighting. The straight line of the weighted regression is markedly displaced downwards due to the large influence of China.

In order to find out which countries exert the strongest potential and actual influence on the regression model, leverage values and Cook's distances were

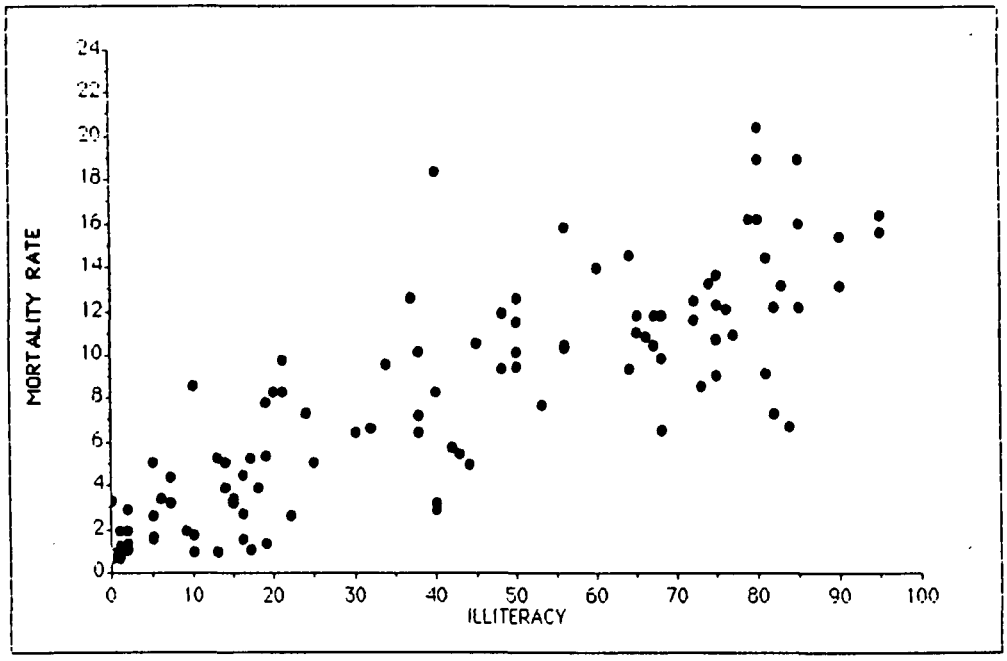


FIGURE 1. *Relation between mortality rates and illiteracy.*

calculated for each country (see earlier). Figure 3 shows the position of each country in the space of the two explanatory variables ILLIT and GDP ('X-space'). In this figure, areas of circles are proportional to leverage values. The countries with the largest leverage values are Saudi Arabia, Oman, United Arab Emirates (UAE), Vietnam and Kuwait. These countries have the strongest potential influence on the regression. As the figure shows the position of these countries is on the border of the 'X-space'.

Analogous to Figure 3, Figure 4 shows the actual influence as measured by Cook's distance as areas of circles in 'X-space'. The countries with the largest Cook's distances are Somalia, Congo, Afghanistan, Iraq, Oman and Saudi Arabia. These countries have the strongest actual influence.

## DISCUSSION

In the following discussion the problems due to the different 'sizes' of the countries are considered first. Subsequently the 'overall' results are described, and the potential and actual influences of the individual countries are discussed. Finally the stability of the identified regression model is investigated under inclusion and exclusion of the countries with the largest actual influence.

The differences in 'size' ie in number of births between the countries are very large. China eg has almost 600 times as many births per year than the UAE. This gave rise to the question of appropriately weighting the countries. The application of Pocock's[9] method to the present data showed that the proportion of residual variance due to sampling is almost zero and that therefore all countries should receive the same weight. This finding contradicts the intuitive impression which suggests that China should receive a much larger weight than eg Bhutan. China and Bhutan provide the same amount of information about the relation between infant mortality and explanatory variables. Their influence on the regression model due to size should therefore be the same.

Performing the appropriate unweighted stepwise regression the following 'overall' results were found: Out of all available explanatory variables the regression selected only two: illiteracy and GDP per capita. The variable illiteracy alone gave a coefficient of determination $R^2 = 0.75$, ie 75% of the variation in infant mortality is 'explained' by illiteracy (Figure 2). The second variable able to enlarge markedly $R^2$ was GDP ($R^2 = 0.82$, partial F-test: $P<0.0001$).

No additional regressor variable (eg number of inhabitants per physician) did increase $R^2$ significantly. It is therefore found that illiteracy plays a major role in 'explaining' infant mortality. In particular, it explains more of the variation than GDP. These results are consistent with the findings of Sagan and Afifi.[10] In their

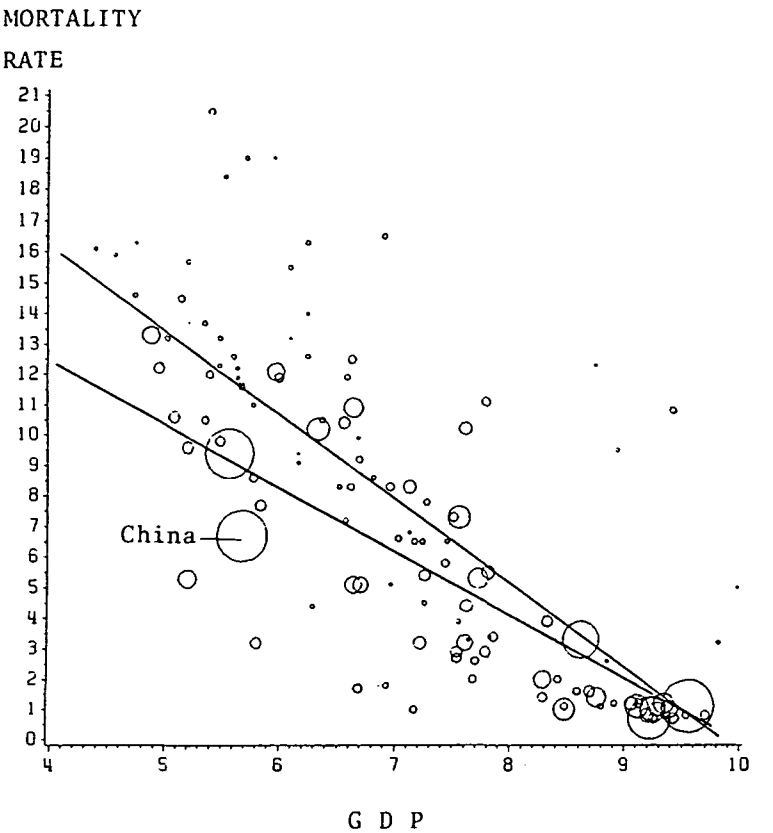


FIGURE 2. *Relation between mortality rates and GDP per capita (logarithmically transformed). Size of the areas proportional to the weights. Upper straight line: without weighting. Lower straight line: with weighting.*

investigation about the relation between infant mortality, energy consumption and other variables they found that illiteracy had the largest coefficient of correlation with infant mortality.

Figure 3 shows the *potential* influence which each country exerts on the model due to the values of the explanatory variables alone, ie to its position in 'X-space'. The areas of the circles are proportional to the potential values of the corresponding countries. One recognizes clearly that the closer a country lies towards the border of the 'X-space', the stronger is its potential influence.

The following countries have the largest 'distance' from the centre in X-space and therefore the largest potential influence: Saudi Arabia, Oman, UAE, Vietnam and Kuwait; four of these are oil-rich countries. As one can see from the figure, they have a common characteristic property: A high degree of illiteracy inspite of the relatively high GDP. Vietnam has the fourth largest potential influence. Its position in 'X-space' is just opposite to the oil-rich countries: Even though the GDP is relatively low, illiteracy is relatively low.

Analogous to Figure 3, Figure 4 shows the *actual* influence of each country. Here the areas of the circles are proportional to Cook's distances. This case statistic combines information from the leverage values and from the residuals. It measures the actual influence of each unit on the regression. Comparing the potential influence of the four oil-rich countries (Figure 3) with their actual influence (Figure 4), one sees that these countries split in two groups. While the actual influence of Saudi Arabia and Oman is strong, the influence of UAE and Kuwait is weak. UAE and Kuwait certainly have large potential values but their residuals are small. This means that even though they are far from the centre in X-space, their observed infant mortality corresponds to what one expects on the basis of the regression model. In contrast to this, Saudi Arabia and Oman also have large Cook's distances. Infant mortality in these countries is larger than one would expect from the model (positive residuals).

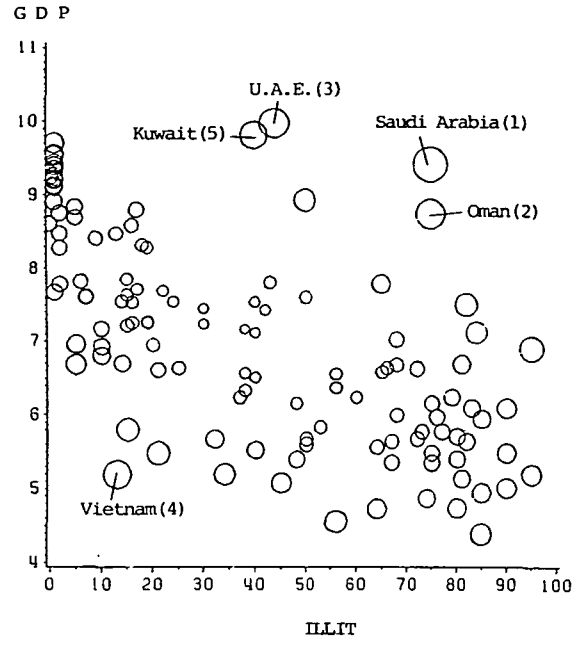


FIGURE 3. *Representation of the potential influence. Size of the areas proportional to potential values. Explanation in the text.*

Somalia has the largest Cook's distance. The observed rate is here larger than expected (positive residual). It is not obvious whether this observed result corresponds to a real effect or unreliable data. The same is found for Afghanistan (positive residual) and Congo and Iraq (negative residuals).

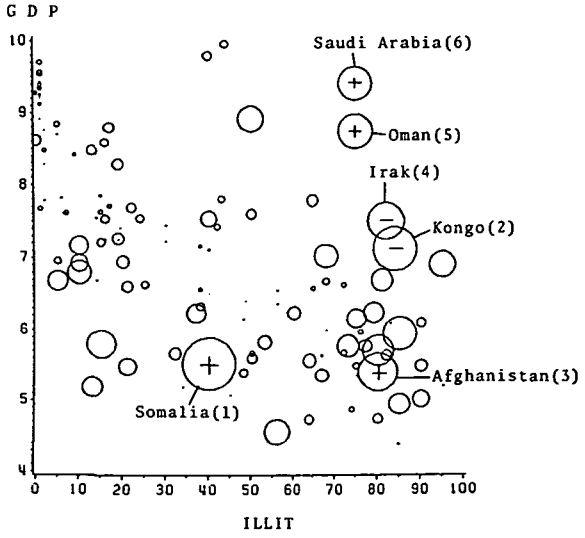


FIGURE 4. *Representation of the actual influence. Size of the areas proportional to Cook's distances. The signs of the residuals are represented by + or − respectively. Explanation in the text.*

In order to find out whether any country exerts an *unduly* large influence, the regression was performed with and without the country with the largest Cook's distance (Somalia). The differences in the estimates of the parameters were very small. The same was found for the other countries with large Cook's distances. This means that the above conclusions about the relation of infant mortality and structural variables are independent of the inclusion or exclusion of 'extreme' countries.

The above example demonstrates that when the observational units are countries, each of which has its own specific structure, it is of particular interest to detect what potential and actual influence each exerts on the regression and with that on the interpretation of the results. Similar applications of measures of influence may arise in many other epidemiological questions. In environmental epidemiology one is interested eg in the relation between respiratory diseases and air pollutants. If the observational units are towns, one can build a regression model which 'explains' the dependent variable in terms of independent variables. However, thereafter it might be equally important for purposes of the population surveillance to go back to the individual units and to make a statement about any particular town, based on its distance from the centre in X-space (potential influence) and on its overall influence on, and its deviations from the model.

REFERENCES

[1] Draper N R, Smith H. *Applied Regression Analysis* (2nd ed), New York: Wiley, 1981; 85–92.

[2] Cook R D. Detection of influential observations in linear regression. *Technometrics* 1977; **19:** 15–8.

[3] Cook R D. Influential observations in linear regression. Detection of influential observations in linear regression. *J Am Stat Assoc* 1979; **74:** 169–74.

[4] Haefs H. *Der Fischer Weltalmanach 1986.* Frankfurt: Fischer Taschenbuch Verlag. 1985.

[5] Woolhandler S, Himmelstein D U. Militarism and mortality, an international analysis of arms spending and infant death rates. *Lancet* 1985; **1:** 1375–8.

[6] Weisberg S. *Applied Linear Regression* (2nd ed), New York: Wiley, 1985; 111–20.

[7] Cook R D, Weisberg S. *Residuals and Influence in Regression.* London: Chapman and Hall, 1982; 115–6.

[8] Fryer J G, *et al.* Comparing the early mortality rates of the local authorities in England and Wales. *J Roy Stat Soc A* 1979; **142:** 181–98.

[9] Pocock S J, *et al.* Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Appl Stat* 1981; **30:** 286–95.

[10] Sagan L A, Afifi A A. *Health and Economic Development I: Infant mortality. RM-78-42.* International Institute for Applied Systems Analysis, Laxenburg, Austria, 1978.

*(Revised version received July 1989)*